Deep learning to segment liver metastases on CT images: Impact on a radiomics method to predict response to chemotherapy

Deep learning to segment liver metastases on CT images: impact on a radiomics method to predict response to chemotherapy

Valentina Giannini
Dept. of Surgical Science
University of Turin and
Candiolo Cancer Institute, FPO-IRCCS
Candiolo (TO), Italy
valentina.giannini@ircc.it

Arianna Defeudis
Dept. of Surgical Science
University of Turin and
Candiolo Cancer Institute, FPO-IRCCS
Candiolo (TO), Italy
arianna.defeudis@ircc.it

Samanta Rosati
Dept. of Electronics and Telecommunications
Polytechnic of Turin
Torino, Italy
samanta.rosati@polito.it

Giovanni Cappello
Dept. Radiology
Candiolo Cancer Institute, FPO-IRCCS
Candiolo (TO), Italy
giovanni.cappello@ircc.it

Lorenzo Vassallo
Dept. Radiology
Candiolo Cancer Institute, FPO-IRCCS
Candiolo (TO), Italy
lorenzo.vassallo@ircc.it

Simone Mazzetti
Dept. of Surgical Science
University of Turin, and
Candiolo Cancer Institute, FPO-IRCCS
Candiolo (TO), Italy
simone.mazzetti@ircc.it

Jovana Panic
Dept. of Surgical Science
University of Turin and
Candiolo Cancer Institute, FPO-IRCCS
Candiolo (TO), Italy
jovana.panic@ircc.it

Daniele Regge
Dept. of Surgical Science
University of Turin, and
Candiolo Cancer Institute, FPO-IRCCS
Candiolo (TO), Italy
daniele.regge@ircc.it

Gabriella Balestra
Dept. of Electronics and Telecommunications
Polytechnic of Turin
Torino, Italy
gabriella.balestra@ircc.it

***Abstract*— Predicting response to neo-adjuvant chemotherapy of liver metastases (mts) using CT images is of key importance to provide personalized treatments. However, manual segmentation of mts should be avoid to develop methods that could be integrated into the clinical practice. The aim of this study is to evaluate if and how much automatic segmentation can affect a radiomics-based method to predict response to neoadjuvant chemotherapy of individual liver mts. To this scope, we developed an automatic deep learning method to segment liver mts, based on the U-net architecture, and we compared the classification results of a classifier fed with manual and automatic masks. In the validation set composed of 39 liver mts, the automatic deep-learning algorithm was able to detect 82% of mts, with a median precision of 67%. Using manual and automatic masks, we obtained the same classification in 19/32 mts. In case of mts with largest diameter > 20 mm, the precision of the segmentation does not impact the classification results and we obtained the same classification with both masks. Conversely, with smaller mts, we showed that a Dice coefficient of at least 0.5 should be obtained to extract the same information from the two segmentations. This are very important results in the perspective of using radiomics-based approach to predict response to therapy into clinical practice. Indeed, either precisely manually segment all lesions or refine them after automatic segmentation is a time-consuming task that cannot be performed on a daily basis.**

Keywords—deep learning, radiomics, automatic segmentation, CT imaging, prediction of response.

I. INTRODUCTION

Colorectal cancer (CRC) is the third most common tumor worldwide [1] and frequently metastasized into thoracic organs (liver and peritoneum). The high heterogeneity that characterizes CRC can lead to differences in response to treatment, either between the primary tumor and the metastatic lesions or among different metastatic lesions in the same patient [2]. Recently, some studies developed radiomics-based algorithms aiming at predicting response to neo-adjuvant chemotherapy of liver metastases (mts) using CT images [3]–[5].

However, all these studies performed the analyses from masks that were manually segmented by a radiologist. Manual segmentation is a time consuming task that is prone to errors and that shows high inter-reader variations [6]. Moreover, in a clinical perspective, where time is an important issue, it would be not feasible to manually segment all mts, before applying computerized methods.

The aim of this study is twofold. First, we want to assess whether automatic segmentation can affect a radiomics-based method to predict response to neoadjuvant chemotherapy (nCT) of individual liver mts. Secondly, we want to evaluate how precise should

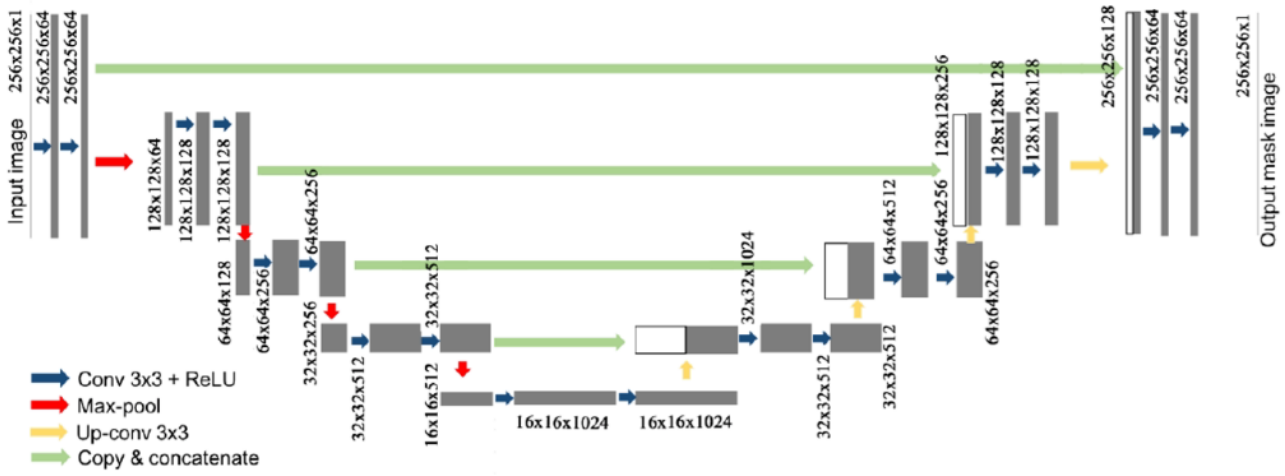

Figure 1: U-net architecture

be the automatic segmentation to not compromise the results of the prediction. To this scope, we developed an automatic deep learning method to segment liver mts, based on the U-net architecture, and we compared the classification results of a classifier fed with manual and automatic masks.

II. MATERIAL AND METHODS

A. Patients and reference standard

We retrospectively included patients with a stage IV CRC having at least one measurable liver mts as defined by the RECIST 1.1 Criteria (greater diameter $\geqslant$ 10 mm). All patients underwent a CT exam with contrast injection within 2 weeks from the start of nCT. A maximum number of 10 liver mts per patient were segmented by a resident radiologist (5 years of experience in reading CT images) on the portal phase of the baseline CT exam using ITK-Snap. All slices of each metastases were manually contoured. Metastases that either were confluent or subdiaphragmatic or contained large vessels or were difficult to measure in the subsequent exams were excluded. Once all mts were segmented, their longest diameter was measured at baseline and after 12 weeks of nCT. Mts were defined as non-responder (R-) if their diameter increased more than 3 mm and responder (R+) if their diameter decreased more than 3 mm or remained stable (±3 mm). This cut-off was chosen based on a preliminary study, in which we demonstrated that 95% confidence interval on the difference between means of diameters of liver mts in CT exams measured by two radiologists was 3 mm. Patients were divided into a training and a validation set. The study was approved by the local Ethics Committee and informed consent was signed by all patients.

B. Automatic segmentation

The automatic segmentation of the liver mts is carried by a U-net based system. The U-net is a Fully Convolutional Network, which provides a prediction mask whose dimensions are the same of the input image.

The network implemented is defined by 4 descending layers (Figure 1). Each of them is characterized by two subsequent Convolutional layers with 3x3 kernel and ReLU activation function [7], and a Max pooling layer, with pool size of 2x2 to halve the dimension of the image. The extending layers are similar to the previous one, where the Max Pooling layer is replaced by the upsampling layer. The output layer is characterized by the Convolutional layer with 1x1 kernel and Softmax activation function. The optimizer is Adam [8] with learning rate of 0.0001, the loss function used is the Binary Crossentropy (1)

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1 - y_i) \cdot \log\big(1 - (y_i)\big), \quad (1)$$

where y_i is the label and $p(y_i)$ is the predicted probability of the sample to belong to the label class. The training epoch was set to 50.

All metastases of patients belonging to the training set were used to develop the U-net. In particular, all slices were divided into a training (700 slices) and a test (300 slices) set. Both sets are composed of the same proportion of healthy and unhealthy slices. The network has been implemented on Matlab R2019a with the Deep Learning toolbox.

Since the number of background pixels is higher than the mts ones, the method of *the inverse frequency weighting* (2) was applied, to balance the provided dataset.

$$weight_{class} = \frac{number\ all\ pixels}{number\ pixels\ belonging\ to\ class} \quad (2)$$

Finally, the U-net was validated using all metastases of patients of the validation set.

C. Classification

In this study, we used a previously developed machine learning algorithm that classifies lesions as R+

TABLE 1: RESULTS OF THE U-NET IN THE VALIDATION SET

ID Pat	*Mean Dice*	*Mean Precision*	*Mean Recall*
1002	0.49	0.41	0.84
1009	0.49	0.64	0.39
1010	0.51	0.41	0.69
1013	0.52	0.60	0.61
2015	0.73	0.75	0.79
3003	0.48	0.48	0.48
3005	0.70	0.66	0.77
3010	0.80	0.72	0.90

and R-, based on radiomics features (RF) computed on the CT scan acquired before nCT [9]. Briefly: a) RFs are extracted from a 7x7 ROI that moves across the image by step of 2 pixels and that is fully included in the tumor mask; b) feature selection is performed using a genetic algorithm (GA); c) classification of each ROI is performed by a two-layer neural network, in which the number of neurons is optimized by the GA during the feature selection step; d) each mts is classified as R+ if the percentage of R+ ROIs is higher than the value represented by the Youden Index computed on the Receiver Operating Characteristics (ROC) curve of the training set. The classifier was trained with ROIs computed using the manually segmented masks of the training set and was subsequently validated on both manual and automatic masks of patients belonging to the validation set.

D. Statistical analysis

Results of segmentation were evaluated on both the training and the validation datasets using 3 metrics: Dice Similarity Coefficient (DSC), Recall and Precision, using the following equations (3)

$$\mathrm{DSC} = \frac{2|\mathrm{MM} \cap \mathrm{AM}|}{|\mathrm{MM}| + |\mathrm{AM}|}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (3),$$

where MM is the manual mask, AM is the automatic mask. TP the number of true positive pixels, FP the number of FP pixels and FN the number of false negative pixels.

DSC computes the overall overlap between two masks, while Precision and Recall are correlated to the under and over segmentation, respectively.

Results of classification were evaluated by computing the confusion matrices and their corresponding values of sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV).

TABLE 2: RESULTS OF THE TRAINING AND VALIDATION SET. NUMBERS ARE IN PERCENTAGES

	Sens (95%CI)	*Spec (95%CI)*	*PPV (95%CI)*	*NPV (95%CI)*
Training	41.3 (29.0-54.4) [26/63]	71.4 (47.8-88.7) [15/21]	81.3 (67.5-90.1) [26/32]	28.9 (22.4-36.3) [15/52]
Validation manual	57.9 (33.5-86.1) [11/18]	61.5 (31.6-86.1) [8/13]	68.8 (50.0-82.9) [11/16]	50.0 (33.6-66.4) [8/16]
Validation U-net	52.6 (28.9- 75.5) [10/19]	61.5 (31.6-86.1) [8/13]	66.7 (47.1-81.8) [10/15]	47.1 (31.9-62.8) [8/17]

TABLE 3: RESULTS OF THE TRAINING SET: NUMBERS OF CORRECTLY CLASSIFIED R+/R- LESIONS OVER THE TOTAL NUMBER OF R+/R- LESIONS.

ID Pat	*# correctly classified R+*	*# correctly classified R-*
1003	2/4	-
1005	3/10	-
1006	0/1	-
1015	1/1	-
1018	1/6	-
1023	3/9	-
1025	1/1	-
2004	1/1	7/7
2010	1/1	-
2011	8/10	-
2012	0/4	-
3004	1/1	-
3006	-	5/6
3007	2/10	-
3009	-	0/2
3012	1/4	4/6

III. RESULTS

A. Patients

The training set was composed of 16 patients for a total of 84 mts (21 R- and 63 R+), while the validation set was composed of 39 mts (22 R+ and 17 R-) from 8 patients. Median size of mts was 28(25^{th}-75^{th} percentile: 19-36) mm.

B. Segmentation

The U-net classified all mts in the training set, with a median Dice overlap of 0.54(25^{th}-75^{th} percentile: 0.30-0.77), a median precision of 0.67(25^{th}-75^{th} percentile: 0.48-0.80), a median recall of 0.66(25^{th}-75^{th} percentile: 0.22-0.82). In the validation set, 7/39 mts were not detected by the automatic segmentation performed by the U-net. Among them, 6 were very small mts (longest diameter≤12), and one had longest diameter=16mm. Mean Dice, recall and precision are shown in Table 1. Examples of segmentations are shown in Figure 2.

C. Classification

In the training set, 15/21 R- and 26/63 R+ mts were correctly classified, leading to a sensitivity and a

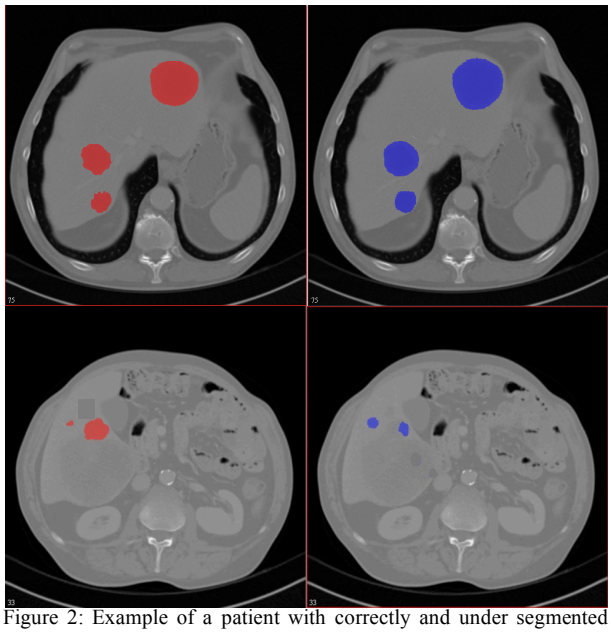

Figure 2: Example of a patient with correctly and under segmented lesions (in red manual masks and in blue automatic masks). First row shows thre metastases (1,2,3) correctly classified (Dice 0.85,0.83,0.71). Second row: one small mts (mts 6) correctly segmented (Dice 0.73) , and one mts (9) under-segmented (Dice=0.17). Both mts are classified differently between manual and automatic masks.

specificity of 41% and 71%, respectively (Table 2). Number of lesions divided per patient is shown in Table 3. 26 out of 37 FNs belonged to only 4 patients (3R+ and 1 R-), moreover 21 out of 26 lesions had the longest diameter smaller than 20 mm. In the validation set 8/13 R- and 11/19 R+ mts were correctly classified, when using the manual mask, and 8/13 R- and 10/19 R+ mts were correctly classified, when using the automatic masks (Table 2). The classification obtained using different masks was congruent in 19 out of 32 mts (Table 4). Table 4 shows that if dice overlap is very low (<0.25) the classification is different regardless the size of the lesion, e.g., the 7th lesion of patient 1013. Indeed, 5/13 lesions with different classification had Dice ≤0.21. Dice overlaps between 0.25 and 0.6 could impact the results if the lesion is very small (longest diameter <15 mm). In our dataset, this is true in 2/13 cases (patient 1009, lesion 2 and patient 2015, lesion 4). High values of Dice (between 0.61 and 0.80) could impact the classification if the lesion is small, while very high Dice (>0.8) does not affect the classification results, independently of lesion diameter.

TABLE 4: DIFFERENCES AMONG CLASSIFICATIONS OBTAIEND WITH MANUAL AND AUTOMATIC MASKS (IN BOLD LESIONS CLASSIFIED DIFFERENTLY).

ID Pat	*ID mts*	*Real class*	*Manual*	*U-net*	*Dice*	*Longest Diameter*
1002	**1**	**1**	**1**	**0**	**0.18**	**30**
1002	2	1	0	0	0.29	28
1002	3	1	0	0	0.64	60
1002	4	1	1	1	0.85	23
1009	1	1	1	1	0.49	10
1009	**2**	**1**	**0**	**1**	**0.48**	**10**
1010	1	1	0	0	0.51	15
1013	1	1	1	1	0.52	43
1013	**2**	**1**	**1**	**0**	**0.73**	**14**
1013	**3**	**1**	**1**	**0**	**0.75**	**14**
1013	4	1	1	1	0.46	21
1013	5	1	1	1	0.68	20
1013	**6**	**1**	**0**	**1**	**0.05**	**17**
1013	**7**	**1**	**0**	**1**	**0.21**	**50**
1013	8	1	1	1	0.82	19
1013	**9**	**1**	**0**	**1**	**0.21**	**38**
1013	**10**	**1**	**1**	**0**	**0.74**	**18**
2015	1	0	1	1	0.68	28
2015	2	0	0	0	0.83	33
2015	3	0	0	0	0.85	26
2015	**4**	**1**	**1**	**0**	**0.55**	**14**
3003	1	0	0	0	0.48	16
3005	1	0	0	0	0.85	59
3005	2	0	1	1	0.83	45
3005	3	1	0	0	0.71	28
3005	4	0	0	0	0.82	31
3005	5	0	0	0	0.69	21
3005	**6**	**0**	**0**	**1**	**0.73**	**11**
3005	**7**	**0**	**1**	**0**	**0.64**	**16**
3005	8	0	1	1	0.82	18
3005	**9**	**0**	**0**	**1**	**0.17**	**31**
3010	**1**	**0**	**1**	**0**	**0.80**	**15**

IV. DISCUSSION

In this study we assessed the impact that an automatic segmentation could have when using a radiomics-based algorithm to predict response to therapy of liver mts. To this scope, we first develop an automatic deep-learning algorithm that was able to detect 82% of liver mts, with a median precision of 67%. High values of precision indicate that the lesion is not over-segmented, which is important in a radiomics based method, since we had to avoid including healthy tissue in the analysis. The precision reached by our method is promising, considered the limited number of cases, and could be improved by including more patients. However, in this study we showed that in case of lesion with diameters ≥ 20 mm (most of mts), the precision of the segmentation does not impact the classification results. Indeed, when Dice overlap is higher than 0.2, we obtain the same classification between manual and automatic masks, meaning that the same information is provided to the classifier. This is a very important results in the perspective of using radiomics-based approach to predict response to therapy into the clinical practice. Indeed, either precisely manually segment all lesions or refine them after automatic segmentation is a time-consuming task that cannot be performed on a daily basis.

Conversely, we demonstrated that in case of very small lesions, it is necessary that the automatic segmentation reaches very high value of Dice overlap to obtain the same classification that we can obtain using manual masks. Indeed, since in this case we have a few

numbers of ROI, we cannot extract the same information from two masks that differ significantly.

One limitation of this study relies on the not optimal results obtained by the classification method. This might be due to the low number of patients included in the study. However, the aim of this study was not to optimize the classifier, but to assess differences between manual and automatic segmentation. Having demonstrated that automatic segmentations, even not perfect, can provide the same information of manual masks in most cases, it would be possible to improve classification method by adding more cases, without relying in manual segmentations.

ACKNOWLEDGMENT

This work was funded by FONDAZIONE AIRC under 5 per Mille 2018 - ID. 21091 program – P.I. Bardelli Alberto, G.L. Regge Daniele.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, 2018, doi: 10.3322/caac.21492.

[2] C. J. A. Punt, M. Koopman, and L. Vermeulen, "From tumour heterogeneity to advances in precision treatment of colorectal cancer," *Nature Reviews Clinical Oncology*. 2017, doi: 10.1038/nrclinonc.2016.171.

[3] S. X. Rao *et al.*, "CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy?," *United Eur. Gastroenterol. J.*, 2016, doi: 10.1177/2050640615601603.

[4] R. C. J. Beckers *et al.*, "CT texture analysis in colorectal liver metastases and the surrounding liver parenchyma and its potential as an imaging biomarker of disease aggressiveness, response and survival," *Eur. J. Radiol.*, 2018, doi: 10.1016/j.ejrad.2018.02.031.

[5] R. Klaassen *et al.*, "Feasibility of CT radiomics to predict treatment response of individual liver metastases in esophagogastric cancer patients," *PLoS One*, 2018, doi: 10.1371/journal.pone.0207362.

[6] R. Meier *et al.*, "Clinical Evaluation of a Fully-automatic Segmentation Method for Longitudinal Brain Tumor Volumetry," *Sci. Rep.*, 2016, doi: 10.1038/srep23376.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, 2017, doi: 10.1145/3065386.

[8] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[9] V. Giannini *et al.*, "An innovative radiomics approach to predict response to chemotherapy of liver metastases based on CT images," in *2020 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'20*, 2020.